FERENCE & ASSOCIATES
PATENT FILING TRANSMITTAL 06-14-00 *DOCKET NO. YOR-2000-0168US1*

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Box Patent Application
Commissioner of Patents and Trademarks
Washington, D.C. 20231

PATENT FILING TRANSMITTAL

Transmitted herewith for filing is the Patent Application of: Upendra V. Chaudhari, Stephane H. Maes, and Jiri Navratil

For: SPEAKER RECOGNITION METHOD BASED ON STRUCTURED SPEAKER MODELING AND A "PICKMAX" SCORING TECHNIQUE

TYPE OF FILING

This new patent application is for a(n):

[X] Utility
[ ] Design
[ ] Plant
[ ] Divisional
[ ] Continuation
[ ] Continuation-in-part

Benefit of a prior filed application

[ ] This application claims the benefit of an earlier filed U.S. Patent Application under 35 USC 120.
[ ] Please accord Applicant the benefit of the priority date of _____to this case pursuant to 35 USC 119. Applicant's claim for priority is based on application _____filed in _____on that date.

Filing under 37 CFR 1.53 (Utility) or 37 CFR 1.153 (Design)

[X] This is an application filed pursuant to 37 CFR 1.53 or 37 CFR 1.153, permitting receipt of a filing date upon filing of a specification, at least one claim and necessary drawings.
[X] In the event any parts of this application are incomplete, please treat this as a filing under 37 CFR 1.53 or 37 CFR 1.153.

ENCLOSURES

[X] 15 - pages of written description;
[X] 10 - pages of claims;
[X] 1 - pages of abstract;
[ ] ___ - sheets of formal drawings;
[X] 3 - sheets of informal drawings;
[X] Declaration and Power of Attorney or listing of inventors;
and
[X] Two postcards for return to us as proof of receipt of the above documents.

plus
[ ] An Assignment of the invention to IBM Corporation and an Assignment cover sheet;
[ ] Verified Statement Claiming Small Entity Status (37 CFR 1.9(f) and 1.27(b))

[ ] Form PTO-1449 (IDS) and two copies of the references listed thereon;
[ ] A certified copy of_____(country) patent application number (priority document).
[ ] A preliminary amendment;
[ ] Declaration of Biological Deposit;
[ ] Submission of sequence listing, computer readable copy and/or amendment relating thereto for biotechnology invention containing nucleotide and/or amino acid sequence;
[ ] An associate power of attorney;
[ ] Other.

DECLARATION OR OATH

The enclosed Declaration or Oath has been executed by:
[ ] Inventor(s);
[ ] Legal representative of the inventors (37 CFR 1.42 or 1.43);
[ ] Joint inventor or person showing proprietary interest on behalf of an inventor who refused to sign or who cannot be reached and this is a petition required by 37 CFR 1.47 and the statement required by 37 CFR 1.47 is attached;
[X] Has not been executed and is enclosed for the purposes of identifying the inventors.

INVENTORSHIP STATEMENT

The inventorship for all the claims in this application is:
[ ] the same;
[ ] not the same and, as an explanation, a statement is/ will be submitted.

LANGUAGE

The application submitted herewith is:
[X] in English;
[ ] in not in English and in terms of 37 CFR 1.52(d) a verified translation is
  [ ] attached
  [ ] not attached.

FEE CALCULATION

The filing fee has been calculated as shown below:

		SMALL ENTITY OR		OTHER THAN A SMALL ENTITY	
BASIC FEE Design Patent		*$155*	$	$310	$
BASIC FEE Utility Patent		$345	$	$690	$690
EXTRA FEES		RATE	FEE	RATE	FEE
TOTAL CLAIMS 27 MINUS 20=	7	x 9=	$0	x18=	$126
INDEP.CLAIMS 3 MINUS 3 =	0	x 39=	$0	x78=	$0
[ ] MULTIPLE DEP.CLAIM		+135=	$	+270=	$
[ ] ASSIGNMENT		+ 40=	$	+40=	$
[ ] RULE 53 SURCHARGE		+ 65=	$	+130=	$
TOTAL			$		$816

FEE PAYMENT

[ ] Attached is Check No.____ in the sum of $____ to cover the filing fee and, if applicable, the assignment fee.

Respectfully submitted,

Dated: June 13, 2000

Stanley D. Ference III
Reg. No. 33,879

FERENCE & ASSOCIATES
129 Oakhurst Road
Pittsburgh, Pennsylvania 15215
(412) 781-7386
(412) 781-8390-Facsimile

06-14-00

A

PATENT

Docket No. YOR9-2000-0168US1

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

JC828 U.S. PTO
09/593275
06/13/00

Applicant(s)	: Upendra V. Chaudhari et al	Group Art: not yet assigned
Serial No.	: not yet assigned	Examiner: not yet assigned
Filed	: herewith	
For	: SPEAKER RECOGNITION METHOD BASED ON STRUCTURED SPEAKER MODELING AND A "PICKMAX" SCORING TECHNIQUE	

EXPRESS MAIL CERTIFICATE

Express Mail Label No. EL503717289US

Date of Deposit 13 June 2000

I hereby certify that the following attached paper(s) or fee:

Patent Application
Written Description
Claims 1-27
Abstract
Drawings (Figs. 1-3)
Declaration and Power of Attorney (unexecuted)
Patent Filing Transmittal
Certificate of Express Mail
Two Return Postcards

are being deposited with the United States Postal Service "Express Mail Post Office to Addressee" service under 37 C.F.R. 1.10 on the date indicated above and is addressed to the Assistant Commissioner for Patents, Washington, D.C. 20231.

Stanley D. Ference III
(Typed or printed name of person mailing paper)

(Signature of person mailing paper(s) or fee)

Mailing Address:

FERENCE & ASSOCIATES
129 Oakhurst Road
Pittsburgh, Pennsylvania 15215
(412) 781-7386
(412) 781-8390-Facsimile

SPEAKER RECOGNITION METHOD BASED ON STRUCTURED SPEAKER MODELING AND A "PICKMAX" SCORING TECHNIQUE

Field of the Invention

The present invention generally relates to score calculation and normalization in a framework of speaker recognition with phonetically structured speaker models.

Background of the Invention

Typically, in speaker recognition systems, a sample of the voice properties of a target speaker is taken and a corresponding voice print model is built. In order to improve system robustness against impostors in a "verification" mode, it is also typical for a large number of non-target speakers (*i.e.*, "background speakers") to be analyzed, pre-stored, and then used to normalize the voice-print likelihood score of the target speakers.

The voice analysis can be conducted at various levels of phonetic detail, ranging from global (phoneme-independent) models to fine phonemic or subphonemic levels. With several such levels in a system, a problem arises as to how to combine scores from different levels. Combining scores from different levels may be important since it may not always be possible to obtain data at the phonemic level. Particularly, while it is recognized that the voice patterns of a speaker vary with phonemes (or sounds), and are thus better

distinguished by models that are created for individual phonemes, it is sometimes the case that the training data will be sparse. In this case, not all of the phoneme models can be created in a robust way (*i.e.*, in terms of statistical robustness) and therefore have to be combined with models created on a higher level of coarseness (or granularity), such as on broad classes of phonemes (vowels, plosives, fricatives etc.) or on phoneme-independent models, whose robustness is higher. Conventionally, this combination is achieved as a linear interpolation of the model scores from individual granularity levels in a method known as the "back-off" method. A discussion of the "back-off" method can be found in F. Jelinek, "Statistical Methods for Speech Recognition" (MIT Press 1998, ISBN 0262100665). However, this method, as well as other conventional methods, have often been found to be inadequate in providing effective speech verification capabilities.

Accordingly, a need has been recognized in connection with providing a system that adequately and effectively combines scores from the individual levels while avoiding other shortcomings and disadvantages associated with conventional arrangements.

Summary of the Invention

The present invention broadly contemplates, in accordance with at least one presently preferred embodiment, the calculation of scores in such a way that the total likelihood is a weighted sum of the likelihood of all phonetic units at all levels of phonetic

granularity (model grains), and that the weights are derived in such a way that the determination of the robustness and significance of the individual model grains is approached with emphasis.

A particular manner of designing these weights on-the-fly is contemplated herein that takes the actual likelihoods of the test utterance into account and allows for determining the level of distinction as well as the phonetic correspondence on-the-fly using a maximum-likelihood criterion for the individual feature vectors. Apart from the improved accuracy, such an arrangement permits a significant reduction in computation during the verification stage since there is no need for explicit phonetic labeling of the test utterance.

It should be understood that the present invention, in broadly contemplating speaker "recognition", encompasses both speaker verification and speaker identification. With regard to "identification", this may be understood as a task of recognizing a previously enrolled speaker based solely on a test utterance (*i.e.*, no additional identity claims are provided, as opposed to verification). The identification result is the recognized speaker's identity (name, number, etc.; as opposed to the binary "accept/reject" result with verification). Typically, for identification, no background population is necessary for normalization. The task is posed as statistical classification problem and typically solved

using a maximum-likelihood classifier. Identification processes contemplated herein address the calculation of the basis likelihood of a frame given a model (just as in the verification mode). Practical applications for identification include automatic user recognition for adaptation. For instance, a speech-enabled application, *e.g.*, a PC-desktop or a personal email assistant over the telephone, can recognize which user is requesting a service without explicitly requiring his/her name or ID.

In one aspect, the present invention provides a method of providing speaker recognition, the method comprising the steps of: providing a model corresponding to a target speaker, the model being resolved into at least one frame and at least one level of phonetic detail; receiving an identity claim; ascertaining whether the identity claim corresponds to the target speaker model; the ascertaining step comprising the steps of determining, for each frame and each level of phonetic detail of the target speaker model, a non-interpolated likelihood value; and resolving the at least one likelihood value to obtain a likelihood score.

In another aspect, the present invention provides an apparatus for of providing speaker recognition, the apparatus comprising: a target speaker model generator for generating a model corresponding to a target speaker, the model being resolved into at least one frame and at least one level of phonetic detail; a receiving arrangement for

receiving an identity claim; a decision arrangement for ascertaining whether the identity claim corresponds to the target speaker model; the decision arrangement being adapted to determine, for each frame and each level of phonetic detail of the target speaker model, a non-interpolated likelihood value; and resolve the at least one likelihood value to obtain a likelihood score.

Furthermore, the present invention provides in another aspect a program storage device readable by machine, tangibly embodying a program of instructions executable by the machine to perform method steps for providing speaker recognition, the method comprising the steps of: providing a model corresponding to a target speaker, the model being resolved into at least one frame and at least one level of phonetic detail; receiving an identity claim; ascertaining whether the identity claim corresponds to the target speaker model; the ascertaining step comprising the steps of determining, for each frame and each level of phonetic detail of the target speaker model, a non-interpolated likelihood value; and resolving the at least one likelihood value to obtain a likelihood score.

For a better understanding of the present invention, together with other and further features and advantages thereof, reference is made to the following description, taken in conjunction with the accompanying drawings, and the scope of the invention will be pointed out in the appended claims.

Brief Description of the Drawings

Figure 1 illustrates an example of a structure speaker model (voice-print) with three levels and a variable number of units on each level.

Figure 2 illustrates a speaker verification system with the "Pickmax" scoring and structure speaker models.

Figure 3 illustrates a speaker identification system using the "Pickmax" scores and a maximum-likelihood classifier.

Description of the Preferred Embodiments

The target as well as the background speaker population (used for cohort-based score normalization) are enrolled into the system by creating their statistical models in the feature space. The enrollment utterances are preferably phonetically structured using a transcription engine or a phonetic labeler (for example, a balistic decoder as described in copending and commonly assigned U.S. Patent Application Serial No. 09/015,150 or forced alignment as described copending and commonly assigned U.S. Patent Application Serial No. 09/519,327).

Based on the labeling information, the data is preferably structured on predefined levels of phonetic detail into units, for instance, global level, phone-class level, and phone

level. It is to be noted, however, that the levels may not necessarily obey a top-down or bottom-up detail hierarchy as in the present example. Corresponding models are then preferably created for each of the units for a given speaker. These so-called structured models represent the speakers' voice-prints, as shown in Figure 1.

Thus, Figure 1 illustrates a structured speaker model 100 that may include statistical models of different "levels" as discussed above, for instance, a global level 102, a phone-class level 104 and a phone level 106. A global level 102 will preferably involve a model created from all feature vectors, a phone-class level 104 may preferably include models created for broad phonemic classes (*e.g.*, vowels, nasals, plosives, fricatives, liquids etc.), while a phone level 106 may preferably include single phones (*e.g.*, "aa", "oh", "n", etc.).

The disclosure now turns to a process of verification in accordance with a presently preferred embodiment of the present invention, as described herebelow and as illustrated in Figure 2.

With regard to a conventional procedure against which at least one presently preferred embodiment of the present invention may be compared, let U denote a test utterance (203) that includes T feature vectors (frames) $u_1, \ldots u_T$, which utterance is to be verified based on a speaker's claimed identity 200c. In this connection, a "claim" refers to

an identification tag (such as an identification number, label, name, etc.) to which a speaker claims to correspond. A claimed identity, then, may be expressed the speaker (for example) as "my name is Jerry," or "my customer number is 1234". The existence of a claim is essential for the verification.

Given a structured model $M(i,j)$ for a speaker with $1 \le i \le L$ levels of detail and with $1 \le j \le K(i)$ units on the i-th level, the score (as log-probability) for the utterance is calculated in each level separately, whereby explicit labeling information is used to identify the corresponding phonetic unit that is to be used on each level:

$$S(U \mid M) = \frac{1}{T} \sum_{i=1}^{L} a_i \bullet \sum_{t=1}^{T} P(u_t \mid M(i, j, (i,t)) \qquad (1)$$

where a_i is an interpolation constant for level I and $j(i,t)$ is the labeling information (unit) for level i and frame t. As examples of labelling information that could be used as $j(i,t)$, one might encounter, for instance, $j(1,1)=1$ and $j(2,1)=4$, meaning that in the time-frame t=1: on level=1 use unit number 1 (which might be for example the only model if the level is the "global" one), and on level=2 use unit number 4 (which might correspond to a class of phonemes such as "fricatives.")

The formula (1) may now be generalized, in terms of weighing, by assigning specific weights to each of the units at each level (i.e. to each grain) as follows:

$$S(U \mid M) = \frac{1}{T} \sum_{i=1}^{L} \sum_{t=1}^{T} b_{i,j(i,t)} \bullet P(u_t \mid M\{i, j(i,t)\}) \qquad (2)$$

with b_{i,j(i,t)} denoting grain-specific weights that satisfy

$$\sum_{i=1}^{L} \sum_{j=1}^{K(i)} b_{ij} = 1 \qquad (3)$$

The weights b may be derived in a way so as to emphasize a) grains whose training data amount was above average, thus, which are expected to be more robust, or b) grains which showed an above-average contribution to the performance measured on some development data set or c) grains that are significant with respect to the current test utterance, all subject to the constraint (3). The latter method is further refined below and an algorithm for determining the weights on-the-fly is described ("pickmax").

In a "pickmax" technique in accordance with an embodiment of the present invention (step 209), the likelihood score *S* for each of the structured models mentioned above is calculated as the average of the likelihoods of the *T* feature vectors which, in turn, are obtained as the maximum likelihoods computed over all units and all levels of the given speaker's structured model ("pickmax"). This corresponds to deriving the weights b_{i,j} in (2) from the likelihood of the actual utterance frame at the time *t* based on all grains, as follows:

$$b_{ij} = 1 \ for \ \{i,j\} = \arg\max_{1 \le i \le L, 1 \le j \le K(i)} P(u_t \mid M\{i,j\})$$

$$b_{i,j} = 0 \ otherwise$$

Since there is only one such maximum (or only one is taken in cases of two or more numerically equal maxima) the constraint (3) is implicitly satisfied.

Thus, the formula (2) can be rewritten as:

$$S(U \mid M) = \frac{1}{T} \sum_{t=1}^{T} \max_{1 \le i \le L, 1 \le j \le K(i)} P(u_t \mid M(i,j)) \qquad (4)$$

It is to be noted that the number of units on each level and the number of levels may vary across speakers, since there might be less data available from certain speakers, entailing the necessity of omitting certain units altogether. The scores calculated in (4) will thus still be appropriate for such inter-speaker inconsistencies in the models. Unlike in equation (1), in equation (4) there is no labeling information and no need for interpolation constants which typically must be obtained from independent data sets and can be a source of "over-training." By this, what is meant is that the additional constant must be determined on some data. If there is not enough data, this constant will be determined in too specific a manner with respect to the training and will not be sufficiently

general. It is to be noted that the score calculation (2) and (4) is suitable for (and, without modifications can be applied to) both the speaker verification and identification problems, as well as other related tasks, such as speaker segmentation and tracking.

The final verification is preferably carried out by calculating the log-likelihood ratio (also step 209), for instance, as

$$L = S(U \mid M) - \frac{1}{C}\sum_{i=1}^{C} S(U \mid BG_i) \qquad (5)$$

where M denotes the target model and BG_i the i-th background (cohort) model. The value of L, or the "final score" or "discriminant", preferably undergoes a threshold decision (at 211) to either accept or reject (213) the hypothesis that the utterance was spoken by the target speaker. Alternatively, the modified log-likelihood ratio test as described in the copending and commonly assigned U.S. patent application entitled "Weight Based Background Discriminant Functions In Authentication Systems " (U. Chaudhari et al.), filed herewith, may be used instead of the log-likelihood equation (5) set forth above.

Thus, with reference to Figure 2, a speaker verification process 200 may employ various background speaker models 200a, which may be constructed similarly to that indicated at 100 in Figure 1 (*i.e.*, with varying levels of phonetic detail). At the same time, a target speaker's voice print may be obtained at 200b. Model parameters 201a, 201b,

corresponding to background speaker models 200a and 200b, respectively, are then preferably input into the step of pickmax-score calculation and the log-likelihood ratio test at 209. Preferably, also serving as input into test 209 are frames 207 resulting from a test utterance 203, preferably with the intermediary step of feature extraction (205). As discussed above, while many suitable methods exist for undertaking such steps, the processes described in U. Chaudhari et al., *supra*, are believed to be particularly appropriate in this context.

A score 211 (L) resulting from test 209 then preferably is input into decision logic (213), with the end result that a (threshold) decision on acceptance or rejection is made at step 215.

It will be appreciated from the foregoing that the technique described hereinabove with reference to Figure 2 creates phonetically structured speaker models. Using the complete unit ensemble provided by the model, a scoring method then assigns the best matching likelihood to each feature vector frame and thus maximizes the resulting model score. This improves the significance of the those models that carry useful information for that particular frame in the verification and thus their "competitiveness" in the final log-likelihood ratio test.

Furthermore, as the score calculation mechanism (verification stage) works on a frame-by-frame basis and picks the maximum likelihood across all phonetic units, there is essentially no need for explicit labeling information during this stage. This may save a considerable amount of computation normally associated with phonetic analysis.

With reference to Figure 3, the identification of a speaker (*i.e.*, determining identity I [indicated at 320]) based on the test utterance 303 as denoted above and involving the score 311 calculated as (2), or specifically (4), can be carried out as a maximum-likelihood classification:

$$I = \arg \max_{1 \le y \le Y} S(U \mid M_y)$$

with Y denoting the total number of speakers enrolled in (i.e. know to) the identification system. Other components of the identification system in Figure 3 that are analogous to components in Figure 2 bear reference numerals that are advanced by 100.

It should be appreciated that the specific task of "identification" can involve recognition methods such as "speaker segmentation" and "speaker tracking". These tasks will preferably use a likelihood score measure for which the generalized score calculation (2) and its preferred form (4) can be applied. A detailed description of these additional tasks can be found in S. Maes, "Conversational Biometrics," (Proc. of the

European Conference on Speech Communication and Technology [EUROSPEECH'99], Budapest, Hungary, 1999).

It should be appreciated that, in contrast to the processes described hereinabove, conventional techniques typically calculate the speaker scores based on either global (phonetically unstructured) models or on different levels of phonetic detail -- in the latter case, however, smoothing techniques, e.g. linear interpolation, between models with the same phonetic distinction (but on different levels of coarseness) are applied, which entails the necessity of phonetic labeling during test as well as the need for tuning interpolation constants using additional development data.

It is to be understood that the present invention, in accordance with at least one presently preferred embodiment, includes a target speaker model generator, a receiving arrangement for receiving an identity claim and a decision arrangement for ascertaining whether the identity claim corresponds to the target speaker model. Together, the target speaker model generator, receiving arrangement and decision arrangement may be implemented on at least one general-purpose computer running suitable software programs. These may also be implemented on at least one Integrated Circuit or part of at least one Integrated Circuit. Thus, it is to be understood that the invention may be implemented in hardware, software, or a combination of both.

If not otherwise stated herein, it is to be assumed that all patents, patent applications, patent publications and other publications (including web-based publications) mentioned and cited herein are hereby fully incorporated by reference herein as if set forth in their entirety herein.

Although illustrative embodiments of the present invention have been described herein with reference to the accompanying drawings, it is to be understood that the invention is not limited to those precise embodiments, and that various other changes and modifications may be affected therein by one skilled in the art without departing from the scope or spirit of the invention.

Claims

What is claimed is:

1. A method of providing speaker recognition, said method comprising the steps of:

providing a model corresponding to a target speaker, the model being resolved into at least one frame and at least one level of phonetic detail;

receiving an identity claim;

ascertaining whether the identity claim corresponds to the target speaker model;

said ascertaining step comprising the steps of:

determining, for each frame and each level of phonetic detail of the target speaker model, a non-interpolated likelihood value; and

resolving the at least one likelihood value to obtain a likelihood score.

2. The method according to Claim 1, wherein, for each frame and each level of phonetic detail, the non-interpolated likelihood value is a maximum likelihood value.

3. The method according to Claim 2, wherein said step of resolving the at least one likelihood value comprises averaging the at least one likelihood value.

4. The method according to Claim 3, wherein the likelihood value is determined via the following general equation:

$$S(U \mid M) = \frac{1}{T} \sum_{i=1}^{L} \sum_{t=1}^{T} b_{i,j(i,t)} \cdot P(u_t \mid M\{i, j(i,t)\}) \ ;$$

wherein b_{i,j(i,t)} corresponds to grain-specific weights that satisfy

$$\sum_{i=1}^{L} \sum_{j=1}^{K(i)} b_{ij} = 1 \ ;$$

and further wherein:

S is the likelihood score;

U is a test utterance, comprising T frames $u_1 \ldots, u_T$;

$M(i,j)$ is a speaker model, with $1 \le i \le L$ levels of detail and with $1 \le j \le K(i)$ units on the i-th level; and

$P(u_t|M(i,j))$ is the probability that a frame u_t corresponds to a speaker model unit j on the i-th level of phonetic detail of the speaker model.

5. The method according to Claim 4, wherein the likelihood score is determined by the following equation:

$$S(U \mid M) = \frac{1}{T} \sum_{t=1}^{T} \max_{1 \le i \le L, 1 \le j \le K(i)} P(u_t \mid M(i,j)) \ .$$

6. The method according to Claim 1, wherein the at least one level of phonetic detail comprises at least one of the following: a global level; a phonemic level and a sub-phonemic level.

7. The method according to Claim 6, wherein the at least one level of phonetic detail comprises all of the following three levels: a global level; a phonemic level and a sub-phonemic level.

8. The method according to Claim 7, wherein said step of providing a model corresponding to a target speaker comprises creating said target speaker model on the basis of training utterances and providing labeling information for each frame.

9. The method according to Claim 1, wherein said ascertaining step further comprises accepting or rejecting the identity claim.

10. The method according to Claim 9, wherein said step of accepting or rejecting comprises comparing a quantity based on the likelihood score to a predetermined threshold value.

11. The method according to Claim 10, further comprising the steps of:

providing at least one model corresponding to at least one background speaker; and

determining the quantity based on the likelihood score via employing the at least one background speaker model.

12. The method according to Claim 11, wherein said step of determining the quantity based on the likelihood comprises determining a log-likelihood ratio based on the likelihood score.

13. The method according to Claim 12, wherein the log-likelihood ratio is determined by the following equation:

$$L = S(U \mid M) - \frac{1}{C}\sum_{i=1}^{C} S(U \mid BG_i);$$

wherein:

L is the log-likelihood ratio;

S is the likelihood score;

M denotes the target speaker model; and

BG_i denotes the i-th background model.

14. An apparatus for of providing speaker recognition, said apparatus comprising:

a target speaker model generator for generating a model corresponding to a target speaker, the model being resolved into at least one frame and at least one level of phonetic detail;

a receiving arrangement for receiving an identity claim;

a decision arrangement for ascertaining whether the identity claim corresponds to the target speaker model;

said decision arrangement being adapted to:

determine, for each frame and each level of phonetic detail of the target speaker model, a non-interpolated likelihood value; and

resolve the at least one likelihood value to obtain a likelihood score.

15. The apparatus according to Claim 14, wherein, for each frame and each level of phonetic detail, the non-interpolated likelihood value is a maximum likelihood value.

16. The apparatus according to Claim 15, wherein said decision arrangement is adapted to resolve the at least one likelihood value via averaging the at least one likelihood value.

17. The apparatus according to Claim 16, wherein the likelihood value is determined via the following general equation:

$$S(U \mid M) = \frac{1}{T} \sum_{i=1}^{L} \sum_{t=1}^{T} b_{i,j(i,t)} \bullet P(u_t \mid M\{i, j(i,t)\}) \ ;$$

wherein b_{i,j(i,t)} corresponds to grain-specific weights that satisfy

$$\sum_{i=1}^{L} \sum_{j=1}^{K(i)} b_{ij} = 1 \ ;$$

and further wherein:

S is the likelihood score;

U is a test utterance, comprising T frames $u_1 \ldots, u_T$;

$M(i,j)$ is a speaker model, with $1 \leq i \leq L$ levels of detail and with $1 \leq j \leq K(i)$ units on the i-th level; and

$P(u_t|M(i,j))$ is the probability that a frame u_t corresponds to a speaker model unit j on the i-th level of phonetic detail of the speaker model.

18. The apparatus according to Claim 17, wherein the likelihood score is determined by the following equation:

$$S(U \mid M) = \frac{1}{T} \sum_{t=1}^{T} \max_{1 \leq i \leq L, 1 \leq j \leq K(i)} P(u_t \mid M(i,j)) \ .$$

19. The apparatus according to Claim 14, wherein the at least one level of phonetic detail comprises at least one of the following: a global level; a phonemic level and a sub-phonemic level.

20. The apparatus according to Claim 19, wherein the at least one level of phonetic detail comprises all of the following three levels: a global level; a phonemic level and a sub-phonemic level.

21. The apparatus according to Claim 20, wherein said target speaker model generator is adapted to generate said target speaker model on the basis of training utterances and providing labeling information for each frame.

22. The apparatus according to Claim 14, wherein said decision arrangement is further adapted to accept or reject the identity claim.

23. The apparatus according to Claim 22, wherein said decision arrangement is adapted to accept or reject the identity claim via comparing a quantity based on the likelihood score to a predetermined threshold value.

24. The apparatus according to Claim 23, further comprising:

a background speaker model generator for providing at least one model corresponding to at least one background speaker;

said decision arrangement being adapted to determine the quantity based on the likelihood score via employing the at least one background speaker model.

25. The apparatus according to Claim 24, wherein said decision arrangement is adapted to determine the quantity based on the likelihood via determining a log-likelihood ratio based on the likelihood score.

26. The apparatus according to Claim 25, wherein the log-likelihood ratio is determined by the following equation:

$$L = S(U \mid M) - \frac{1}{C}\sum_{i=1}^{C} S(U \mid BG_i);$$

wherein:

L is the log-likelihood ratio;

S is the likelihood score;

M denotes the target speaker model; and

BG_i denotes the i-th background model.

27. A program storage device readable by machine, tangibly embodying a program of instructions executable by the machine to perform method steps for providing speaker recognition, said method comprising the steps of:

providing a model corresponding to a target speaker, the model being resolved into at least one frame and at least one level of phonetic detail;

receiving an identity claim;

ascertaining whether the identity claim corresponds to the target speaker model;

said ascertaining step comprising the steps of:

determining, for each frame and each level of phonetic detail of the target speaker model, a non-interpolated likelihood value; and

resolving the at least one likelihood value to obtain a likelihood score.

SPEAKER RECOGNITION METHOD BASED ON STRUCTURED SPEAKER MODELING AND A "PICKMAX" SCORING TECHNIQUE

Abstract of the Disclosure

A technique for improved score calculation and normalization in a framework of recognition with phonetically structured speaker models. The technique involves determining, for each frame and each level of phonetic detail of a target speaker model, a non-interpolated likelihood value, and then resolving the at least one likelihood value to obtain a likelihood score.

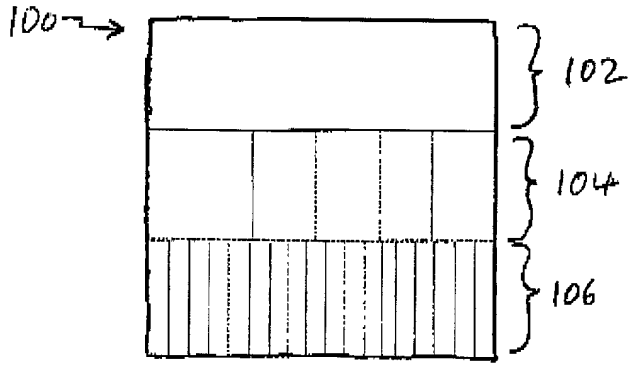

FIG. 1

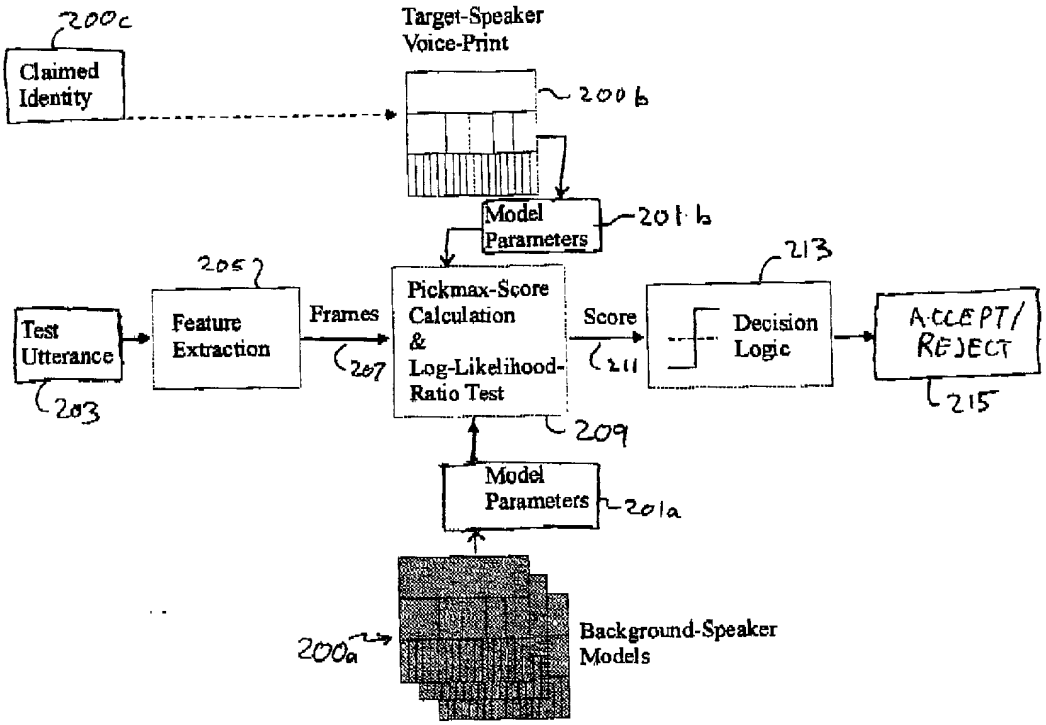
200c
Claimed Identity
Target-Speaker Voice-Print
200b
Model Parameters
201b
205
Test Utterance
203
Feature Extraction
Frames
207
Pickmax-Score Calculation & Log-Likelihood-Ratio Test
209
Score
211
213
Decision Logic
ACCEPT/ REJECT
215
Model Parameters
201a
200a
Background-Speaker Models

FIG. 2

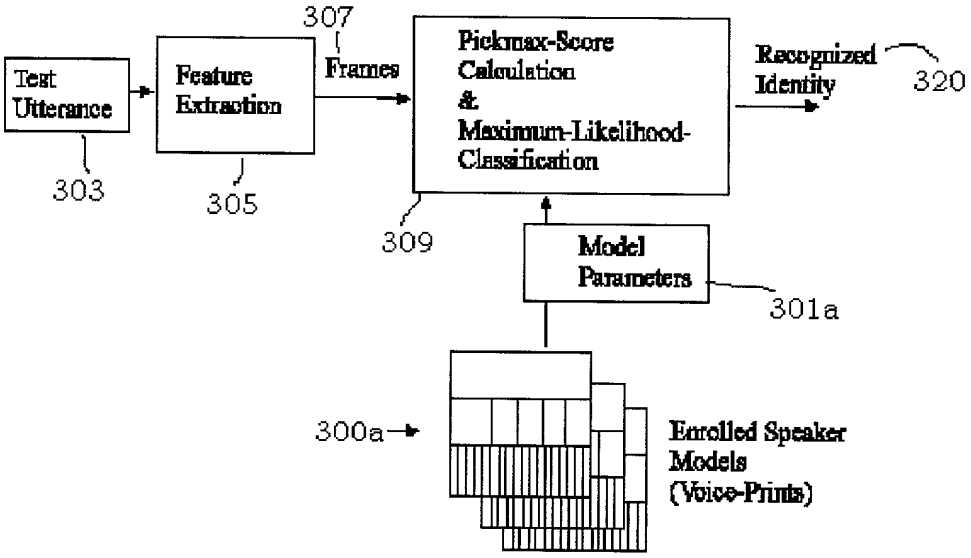
Test Utterance
303
Feature Extraction
305
307
Frames
Pickmax-Score Calculation & Maximum-Likelihood-Classification
309
Recognized Identity
320
Model Parameters
301a
300a
Enrolled Speaker Models (Voice-Prints)

FIG. 3

DECLARATION AND POWER OF ATTORNEY FOR PATENT APPLICATION

As a below named inventor, I hereby declare that:

My residence, post office address and citizenship are as stated below next to my name;

I believe I am the original, first and sole inventor (if only one name is listed below) or an original, first and joint inventor (if plural names are listed below) of the subject matter which is claimed and for which a patent is sought on the invention entitled:

SPEAKER RECOGNITION METHOD BASED ON STRUCTURED SPEAKER MODELING AND A "PICKMAX" SCORING TECHNIQUE

the specification of which (check one)

____ is attached hereto.

__X__ was filed on 13 June 2000 as International Business Machines Docket No. YOR9-2000-0168US1

and was amended on _____ (if applicable)

I hereby state that I have reviewed and understand the contents of the above identified specification, including the claims, as amended by any amendment referred to above.

☐ I acknowledge the duty to disclose information which is material to the patentability of this application in accordance with Title 37, Code of Federal Regulations, Section 1.56.

☐ I hereby claim foreign priority benefits under Title 35, United States Code, §119(a)-(d) or §365(b) of any foreign application(s) for patent or inventor's certificate, or §365(a) of any PCT International application which designated at least one country other than the United States, listed below and have also identified below, by checking the box, any foreign application for patent or inventor's certificate, or PCT International application, having a filing date before that of the application on which priority is claimed:

Prior Foreign Application(s)			Priority Claimed
(Number)	(Country)	(Day/Month/Year Filed)	___ Yes ___ No
(Number)	(Country)	(Day/Month/Year Filed)	___ Yes ___ No
(Number)	(Country)	(Day/Month/Year Filed)	___ Yes ___ No

I hereby claim the benefit under 35 U.S.C. §119(e) of any United States provisional application(s) listed below.

(Application Number)	(Filing Date)
(Application Number)	(Filing Date)

DELARATION AND POWER OF ATTORNEY FOR PATENT APPLICATION

I hereby claim the benefit under 35 U.S.C. §120 of any United States Application(s), or §365(c) of any PCT International application designating the United States, listed below and, insofar as the subject matter of each of the claims of this application is not disclosed in the prior United States, or PCT International application in the manner provided by the first paragraph of 35 U.S.C. §112, I acknowledge the duty to disclose information material to the patentability of this application as defined in 37 CFR §1.56 which occurred between the filing date of the prior application and the national or PCT international filing date of this application:

(Application Serial No.)	(Filing Date)	(Status) (patented, pending, abandoned)
(Application Serial No.)	(Filing Date)	(Status) (patented, pending, abandoned)

I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that willful false statements may jeopardize the validity of the application or any patent issued thereon.

POWER OF ATTORNEY: As a named inventor I hereby appoint the following attorney(s) and/or agent(s) to prosecute this application and transact all business in the Patent and Trademark Office connected therewith (list name and registration number).

Manny W. Schecter (Reg. 31,722), Terry J. Ilardi (Reg. 29,936), Christopher A. Hughes (Reg. 26,914), Edward A. Pennington (Reg. 32,588), John E. Hoel (Reg. 26,279), Joseph C. Redmond, Jr. (Reg. 18,753), Paul J. Otterstedt (Reg. 37,411), Douglas W. Cameron (Reg. 31,596), Wayne L. Ellenbogen (Reg. No. 43,602), Stephen C. Kaufman (Reg. 29,551), Daniel P. Morris (Reg. 32,053), Louis J. Percello (Reg. 33,206), Jay P. Sbrollini (Reg. 36,266), Robert M. Trepp (Reg. 25,933), David M. Shofi (Reg. 39,835, and Louis P. Herzberg (Reg. 41,500)

Send Correspondence to: FERENCE & ASSOCIATES, 129 Oakhurst Road, Pittsburgh, PA 15215

Direct Telephone Calls to: (name and telephone number) Stanley D. Ference III, (412) 781-7386

Upendra V. Chaudhari
Full name of sole or first inventor

Inventor's Signature Date

202 Nob Hill Drive, Elmsford, NY 10523
Residence

USA
Citizenship

Same as above
Post Office Address

DELARATION AND POWER OF ATTORNEY FOR PATENT APPLICATION

Stephane H. Maes
Full name of second joint-inventor, if any

Inventor's Signature Date

1 Wintergreen Hill Road, Danbury, CT 06811
Residence

Belgium
Citizenship

Same as above
Post Office Address

Jiri Navratil
Full name of third joint-inventor, if any

Inventor's Signature Date

154A North Broadway, 2D, White Plains, NY 10603
Residence

Czech Republic
Citizenship

Same as above
Post Office Address